

BẢO HỘ CƠ SỞ DỮ LIỆU LỚN

TS Nguyễn Như Hà

Trường Đại học Khoa học xã hội và Nhân văn, Đại học Quốc gia Hà Nội

Trong văn bản tóm tắt Hiệp định đối tác xuyên Thái Bình Dương (TPP), tại chương 14 về thương mại điện tử, có nêu rõ: “Các nước TPP cam kết bảo đảm luồng thông tin và dữ liệu mang tính toàn cầu được lưu hành một cách tự do giúp phát triển nền kinh tế Internet và kỹ thuật số...”. Đây thực sự là một trong những cơ hội và thách thức to lớn đối với việc phát triển thương mại điện tử ở Việt Nam, bởi lẽ, dữ liệu chính là “nhiên liệu” của nền kinh tế thông tin, là nền tảng cơ bản của các mô hình kinh doanh mới. Tuy nhiên cho tới nay có thể nói, chúng ta chưa thực sự quan tâm tới vấn đề bảo hộ sở hữu trí tuệ đối với sưu tập dữ liệu. Bài viết đề cập đến một số vấn đề cơ bản của dữ liệu lớn (big data) cũng như đề xuất hướng tiếp cận bảo hộ sở hữu trí tuệ đối với sưu tập dữ liệu theo cách thức mà Liên minh châu Âu (EU) đang tiến hành.

Dữ liệu lớn và ứng dụng

Thay vì cách làm phổ biến là hầu hết mọi dữ liệu đều được lưu trữ trong ổ cứng của máy tính, đĩa CD-ROM, USB, thì hiện nay với sự hỗ trợ của nền tảng điện toán đám mây và Internet, mọi thứ đều được số hoá và đưa lên điện toán đám mây để khai thác qua các ứng dụng trên thiết bị thông minh khiến cho thế giới tràn ngập thông tin, sinh ra một lượng dữ liệu khổng lồ¹. Sự thay đổi về quy mô đã dẫn đến một sự thay đổi trạng thái, thay đổi về lượng đã dẫn tới thay đổi về chất. Các khoa học như thiên văn, gen bùng nổ trong những năm 2000 đã đưa ra thuật ngữ “Dữ liệu lớn” và khái niệm này đã làm thay đổi to lớn cách vận hành truyền thống của xã hội, di trú vào tất cả các lĩnh vực của đời sống con người.

Mô hình “5V” - năm tính chất quan trọng hiện đang được thế giới nói đến như những đặc trưng cơ bản của dữ liệu lớn², cụ thể là:

- Volume (số lượng lưu trữ): dữ liệu lớn là tập hợp dữ liệu có dung lượng lưu trữ vượt mức đảm đương của những ứng dụng và công cụ truyền thống. Dữ liệu lớn đang từng ngày tăng lên và tính đến nay nó có thể nằm trong khoảng vài chục terabyte cho đến nhiều petabyte (1 petabyte = 1.024 terabyte)³ chỉ cho một tập hợp dữ liệu mà thôi⁴.

- Velocity (tốc độ xử lý): dung lượng gia tăng của dữ liệu rất nhanh và tốc độ xử lý đang đạt tới mức độ thời gian thực. Các ứng dụng phổ biến trên lĩnh vực Internet, tài chính, ngân hàng, hàng không, quân sự, y tế - sức khỏe ngày hôm nay phần lớn dữ liệu lớn được xử lý trong thời gian thực, xử lý tức thì trước khi chúng được lưu trữ vào cơ sở dữ liệu.

- Variety (đa dạng chủng loại): hình thức lưu trữ và chủng loại dữ liệu ngày một đa dạng hơn. Trước đây chúng ta hay nói đến dữ liệu có cấu trúc thì ngày nay hơn 80% dữ liệu trên thế giới được sinh ra là phi cấu trúc (tài liệu, blog, hình ảnh, video, voice...).

¹Theo Eric Schmidt - CEO của Google, năm 2003 thế giới tạo ra 5 Exabyte dữ liệu (5 tỷ Gigabyte) thì đến năm 2010, cứ 2 ngày thế giới lại tạo ra 5 Exabyte dữ liệu, ước tính năm 2014, cứ 10 phút thế giới lại tạo ra chừng đó dữ liệu. Xem thêm tại địa chỉ: <http://www.bloomberg.com/news/articles/2013-05-30/googles-eric-schmidt-invests-in-obamas-big-data-brains>.

²Xin lưu ý rằng, nhiều nghiên cứu về dữ liệu lớn trên thế giới hiện nay đã xem xét tới mô hình 12V... Tuy nhiên, trong mọi trường hợp, chúng đều xuất phát từ mô hình cơ bản khởi đầu là 3V (volume, variety, velocity).

³1 pétaoctets (PB) = 220 gigaoctets (GB) ~ 1.000.000 GB; 1 téraoctets (TB) = 210 gigaoctets (GB) ~ 1.000 GB.

⁴Xem thêm: LAURILA Juha K., GATICA-PEREZ Daniel, AAD Imad, BLOM Jan, BORNET Olivier, DO Trinh-Minh-Tri, DOUSSE Olivier, EBERLE Julien, Markus MIETTINEN, The mobile data challenge: Dữ liệu lớn for mobile computing research, Workshop on the Nokia Mobile Data Challenge, Proceedings of the Conjunction with the 10th International Conference on Pervasive Computing, 2012, p.1-8.

- Veracity (độ chính xác): một trong những tính chất phức tạp⁵ nhất của dữ liệu lớn là độ chính xác của dữ liệu. Bài toán phân tích và loại bỏ dữ liệu thiếu chính xác và nhiễu là một đặc tính vô cùng quan trọng của dữ liệu lớn.

- Value (giá trị thông tin): giá trị thông tin là tính chất quan trọng nhất của xu hướng công nghệ dữ liệu lớn. Ở đây chúng ta phải hoạch định được những giá trị thông tin hữu ích của dữ liệu lớn cho các vấn đề, bài toán hoặc mô hình hoạt động mà ta hướng tới. Dữ liệu lớn không chỉ đơn thuần là vấn đề kích cỡ và dung lượng của dữ liệu, mà người dùng phải tiếp cận, chọn lọc nguồn dữ liệu, cung cấp thuật toán tối ưu để giúp máy tính có thể phân tích, xử lý và khai thác thông tin nhằm phục vụ cho mục đích của con người.

Dữ liệu lớn và các công nghệ phân tích có khả năng làm thay đổi hoàn toàn bộ mặt của các ngành kinh tế và nghề nghiệp, làm thay đổi cách làm việc và tư duy trong việc khai thác và sử dụng thông tin trong các hoạt động kinh doanh, nghiên cứu. Thời đại dữ liệu lớn đã hiện diện và chi phối đời sống, cách thức tư duy của nhân loại và trong tương lai chắc chắn sẽ tạo ra một trật tự thế giới hoàn toàn khác với hiện nay.

Trên thế giới, có thể kể tên một số trường hợp sau đây, đang ứng dụng rất thành công công nghệ dữ liệu lớn như: Internet (Google, Facebook, Twitter, Amazon, eBay), Mobile (Nokia-MS, Uber app), chăm sóc sức khỏe (IBM Watson, Google Flu), giao thông - vận tải (Fedex, Boeing, F1), điện toán đám mây (Amazon Web Service), chính trị (Digital President Obama 2012), tài chính (New Stock Exchange, JP Morgan), thể thao (English Premier League, Manchester United FC)... Tại Việt Nam, hiện tại có các doanh nghiệp đi tiên phong trong việc áp dụng dữ liệu lớn như: Internet (VCCorp, FPT, VNG), viễn thông (Viettel, FPT Telecom), đa phương tiện (VnExpress, Ren cliquant.vn), giao thông - vận tải (Vietnam Airline)...

Dữ liệu lớn và các ứng dụng có liên quan đang ngày càng được sử dụng rộng rãi trong các cơ quan, tổ chức, trong các lĩnh vực khác nhau của đời sống kinh tế - xã hội. Các phân tích trên lượng dữ liệu lớn còn góp phần cải tiến và tối ưu hóa quá trình ra quyết định, giảm thiểu

rủi ro, tạo ra những giá trị gia tăng cho doanh nghiệp, hỗ trợ tổ chức trong việc quản lý các hoạt động hàng ngày cũng như ra quyết định.

Ngoài ra, cùng với sự phát triển của Internet, của web 2.0, web 3.0, các thiết bị di động cho phép việc sử dụng nhiều phương thức khác nhau để tương tác với khách hàng bên cạnh các phương tiện truyền thống. Việc phân tích các giao dịch của khách hàng qua các kênh khác nhau này giúp doanh nghiệp hiểu hành vi khách hàng, phân cụm nhóm khách hàng, từ đó có thể cung cấp các sản phẩm và dịch vụ phù hợp với yêu cầu khách hàng. Dữ liệu lớn còn mang lại lợi ích cho các doanh nghiệp trong việc lên kế hoạch bán hàng. Bằng việc so sánh các yếu tố khác nhau từ nguồn dữ liệu khổng lồ, doanh nghiệp có thể tối ưu hóa việc định giá cho các sản phẩm. Việc sử dụng dữ liệu lớn trong quản lý chuỗi cung ứng cho phép các doanh nghiệp tối ưu hóa dự trữ kho, vận chuyển, phối hợp với nhà cung cấp nhằm giảm thiểu khoảng cách giữa cung và cầu, kiểm soát ngân sách, cải thiện dịch vụ...

Bảo hộ sưu tập dữ liệu

Bên cạnh các cơ hội mà dữ liệu lớn mang lại, việc tham gia thương mại quốc tế của Việt Nam trong thời gian tới sẽ gặp những thách thức vô cùng to lớn khi mà các nguồn dữ liệu sẽ được tập hợp và khai thác triệt để bởi nhiều chủ thể khác nhau mà chắc chắn chúng ta sẽ phải chịu nhiều thiệt thòi nếu không quan tâm đúng mức đến việc bảo hộ pháp lý đối với dữ liệu. Có thể nói, giống như chương trình máy tính, bảo hộ sưu tập dữ liệu là một vấn đề cực kỳ phức tạp và mang theo mình những rắc rối, đặc thù riêng về mặt pháp lý.

Các quy định quốc tế về bảo hộ sưu tập dữ liệu

Hiện nay, trên thế giới có 3 văn bản chính có các quy định liên quan đến bảo hộ sưu tập dữ liệu, cụ thể là: Công ước Berne về bảo hộ các tác phẩm văn học và nghệ thuật, Hiệp định về các khía cạnh liên quan đến thương mại của quyền sở hữu trí tuệ (TRIPS) và Hiệp ước của WIPO về quyền tác giả (WCT).

Công ước Berne: tại Điều 2(5) quy định: Các tuyển tập các tác phẩm văn học nghệ thuật, các bộ bách khoa từ điển và các hợp tuyển mà do việc chọn lọc hay kết cấu các tư liệu, tạo thành một sáng tạo trí tuệ, cũng được bảo hộ như một tác phẩm, miễn không phương hại quyền tác giả của các tác phẩm tạo nên các hợp tuyển này.

TRIPS: tại Điều 10(2) quy định: Các bộ sưu tập dữ liệu hoặc tư liệu khác, dù dưới dạng đọc được bằng máy hay dưới dạng khác, mà việc tuyển chọn hoặc sắp xếp nội dung chính là thành quả của hoạt động trí tuệ đều

⁵Sự tổ hợp các nguồn dữ liệu này hết sức phức tạp dù cho kích thước của chúng là không lớn. Chẳng hạn một máy bay với một trăm nghìn bộ cảm biến trong một giờ bay chỉ tạo ra 3 gigabytes dữ liệu (100.000 bộ cảm biến x 60 phút x 60 giây x 8 bytes = 3Gb), nhưng chính độ phức tạp của tổ hợp dữ liệu từ các bộ cảm biến này tạo ra tình huống "dữ liệu lớn mà không to". Ngược lại trong nhiều tình huống lượng dữ liệu được sinh ra đều đặn và rất lớn về kích thước, nhưng nếu các dữ liệu này có cấu trúc đơn giản, có quy luật, thì đây lại là tình huống của "dữ liệu to mà không lớn".